Modeling Annotator Subjectivity for Sexism Detection on Social Media

Notebook for the EXIST Lab at CLEF 2025

Qizhang Chen, Leilei Kong*, Yanfu Chen and Changxin Sun

Foshan University, Foshan, China

Abstract

This paper presents Mumul03's submission to the EXIST 2025 shared task on sexism detection. We employ ModernBERT-large, used in this task for the first time, and incorporate demographic information from the annotator such as gender, ethnicity, age, and other attributes into the model input. By modeling individual annotator perspectives and aggregating predictions across submodels, our system effectively captures subjectivity in the annotation process. Our system achieved a ranking of 7th out of 64 submissions for Task 1 in the Soft-Soft category setting. This paper reports our findings on the classification of sexism within textual content on social media, offering substantial insights for the EXIST 2025 challenge.

Keywords

Sexism detection, Text classification, Transformers, EXIST 2025, Natural language processing

1. Introduction

Sexism, defined as prejudice or discrimination based on sex or gender, often targets women and girls through subtle and explicit expressions in everyday life [1, 2]. Although public awareness has grown, this bias remains deeply rooted and is increasingly manifest in online platforms [3]. Online sexism not only undermines women's self-perception and opportunities but also poses risks to young audiences by normalizing misogynistic content. Recent studies show that platforms such as TikTok can quickly expose users to such harmful material [4].

In this context, NLP technologies have become vital tools in detecting and mitigating online sexism [5]. We participate in the EXIST 2025 shared task to explore robust modeling approaches for this challenge.

As part of the EXIST 2025 evaluation, Task 1 is defined as a binary classification task that determines whether a tweet—originally posted on Twitter (now rebranded as X)—contains sexist content, including both explicit and implicit expressions of gender bias. Task 2 is a three-class classification task focused on identifying the author's intent, categorizing tweets into direct sexism, reported sexism, or judgmental commentary [1]. These tasks are designed to improve automatic systems' understanding of both the presence and contextual framing of sexism on social media.

To effectively address these tasks, it is essential to select a model capable of understanding subtle linguistic cues. ModernBERT-large offers an enhanced bidirectional attention mechanism and context-aware encoding, enabling it to capture nuanced semantic signals such as sarcasm and implicit bias [6]. These features are crucial for detecting sexism, where meaning is often shaped by context and subjectivity.

However, relying on a single model to assess whether a statement is sexist introduces bias, as such judgments are inherently subjective. Prior research shows that annotator backgrounds such as gender and cultural context significantly influence their perceptions [7]. Therefore, employing a

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

qizhangc13@gmail.com (Q. Chen); kongleilei1979@gmail.com (L. Kong); c2735066426@163.com (Y. Chen); 18779866976@163.com (C. Sun)

0009-0006-0362-4267 (Q. Chen); 0000-0002-4636-3507 (L. Kong); 0009-0007-2332-0671 (Y. Chen); 0009-0007-9302-7905 (C. Sun)

single predictive model risks masking these individual-level variations, ultimately undermining the representational richness afforded by diverse annotator perspectives [8].

To address this, we adopt the Learning with Disagreement (LwD) paradigm [1] and propose a multi-model ensemble framework built on ModernBERT-large. Each submodel represents a distinct annotator perspective, and we aggregate their predictions to simulate a diversity of viewpoints. This approach better reflects the pluralistic nature of human judgment in the detection of sexism.

The rest of this paper is organized as follows. Section 2 reviews related work; Section 3 details our system; Section 4 presents experiments and results; Section 5 concludes.

2. Relate Work

Recent advances in sexism detection have generally followed three strategic directions: multiview ensembles, which aggregate predictions from multiple models or perspectives to improve robustness; transformer fine-tuning, where large pretrained models are adapted to task-specific data using techniques like data augmentation; and prompt-based large language models (LLMs), which use zero- or few-shot in-context learning to perform sexism detection without task-specific fine-tuning. Each of these approaches has shown strong empirical performance in benchmark evaluations such as EXIST[1] and SemEval [9].

2.1. Multiview Ensembles

Model ensembling enhances robustness, particularly in multilingual and cross-domain contexts. Methods such as prediction averaging have been shown to improve F1 scores and reduce classification uncertainty [10]. At EXIST 2024, the NYCU-NLP team implemented a multilingual, multitask architecture that combined DeBERTa-v3 and XLM-RoBERTa with demographic features of annotators, achieving top performance across all subtasks [11]. Similarly, the CIMAT-CS-NLP team integrated GPT-4-style models with fine-tuned multilingual transformers, ranking among the top systems in Task 1 [12]. The BAZI team optimized XLM-RoBERTa for soft-label uncertainty and secured second place in Task 2 [13].

2.2. Transformer Fine-tuning

Transformer-based encoders are widely used for supervised sexism detection due to their contextual understanding and multilingual adaptability. Commonly used models include XLM-RoBERTa, DeBERTaV3, mBERT, and Spanish-specific variants like BETO and BERTIN. To improve generalization, teams employed techniques such as soft-label learning, multi-label classification, and data augmentation. The AlexPUPB team, for example, fine-tuned compact models like XLM-RoBERTa and MiniLMv2 using soft labels derived from annotator distributions [14]. Back-translation was also used to augment data by introducing lexical variation without altering semantics [15].

2.3. Prompt-based Large Language Models

Prompt-based large language models (LLMs) have emerged as strong baselines in recent EXIST tasks [16, 17]. The mc-mistral team showed that a single Mistral-7B model, prompted with a compact few-shot format mixing English and Spanish examples, outperformed many supervised baselines without fine-tuning [18]. The CIMAT-CS-NLP team achieved competitive zero-shot performance with Google Gemini by combining formal definitions of sexism and expert-role prompts [19]. These results underscore the potential of careful prompt design for low-resource, multilingual classification.

Recent studies have extended prompting by incorporating socio-demographic context. Magnossão de Paula et al. prompted LLMs with demographic personas (e.g., “a male over 45”) and found that model outputs tended to align more with female annotators by default, though alignment effects were inconsistent [20]. Jiang et al. further demonstrated that including annotator profiles—such as gender and ideological stance—improved the model’s ability to replicate individual judgments [21]. These

approaches highlight both the promise and limitations of demographic-aware prompting for subjective NLP tasks.

3. Datasets and Evaluation Metrics

3.1. Datasets

In this study, we only participated in the English subset of Task 1 (sexism identification) and Task 2 (source intention classification) under the soft evaluation setting. Therefore, this section focuses solely on the portion of the dataset relevant to our experiments.

The English subset of the EXIST 2025 dataset consists of 4,727 tweets, which are annotated for various types of sexist expressions, including both explicit and reported sexism. The dataset is divided into a training set (3,260 tweets), a development set (489 tweets), and a test set (978 tweets).

For each sample, the following attributes are provided in a JSON format:

- `id_EXIST`: a unique identifier for the tweet.
- `tweet`: the text of the tweet.
- `number_annotators`: the number of persons that have annotated the tweet.
- `annotators`: a unique identifier for each of the annotators.
- `gender_annotators`: the gender of the different annotators ("F" or "M", for female and male respectively).
- `age_annotators`: the age group of the different annotators (grouped in "18–22", "23–45", or "46+").
- `labels_task1`: a set of labels (one for each of the annotators) that indicate if the tweet contains sexist expressions or refers to sexist behaviors or not ("YES" or "NO").
- `labels_task2`: a set of labels (one for each of the annotators) recording the intention of the person who wrote the tweet ("DIRECT", "REPORTED", "JUDGEMENTAL", "–", and "UNKNOWN").

The dataset is annotated by a diverse group of individuals in terms of gender, age, ethnicity, education level, and country of residence. This diversity contributes to the robustness and fairness of the annotations, helping the dataset to capture a wide range of perspectives and reduce potential annotation bias.

3.2. Evaluation Metrics

This evaluation targets systems that output probability distributions over categories instead of single-class predictions. To effectively assess system performance in scenarios involving multiple and potentially conflicting annotations, ICM-Soft (a soft-label extension of the Information Contrast Measure) is adopted as the official evaluation metric. Additionally, results are also reported using its normalized form (ICM-Soft Norm) and Cross Entropy, providing complementary views on prediction quality.

- `ICM-Soft`: a soft-label extension of the ICM metric, designed to compare the predicted probability distribution with the distribution of human annotations. It is particularly suitable for tasks where label disagreement or subjectivity is common, as it evaluates how closely the model's predictions align with the collective opinions of annotators.
- `ICM-Soft Norm`: a normalized version of ICM-Soft, which rescales the raw scores to facilitate comparison across different tasks or systems.
- `Cross Entropy`: a widely used metric in classification tasks that measures the difference between the predicted and true probability distributions. Lower values indicate better alignment with the true distribution.

4. System Description

Our architecture, illustrated in Figure 1, consists of multiple models based on transformers independently trained. Each model receives a version of the same tweet augmented with the demographic information of a specific annotator. These models are trained using a unified configuration and are designed to capture diverse annotator perspectives. Their predictions are then aggregated to produce the final output. The following subsections provide a detailed breakdown of our system's construction, including data preprocessing, metadata integration, model fine-tuning, and output aggregation.

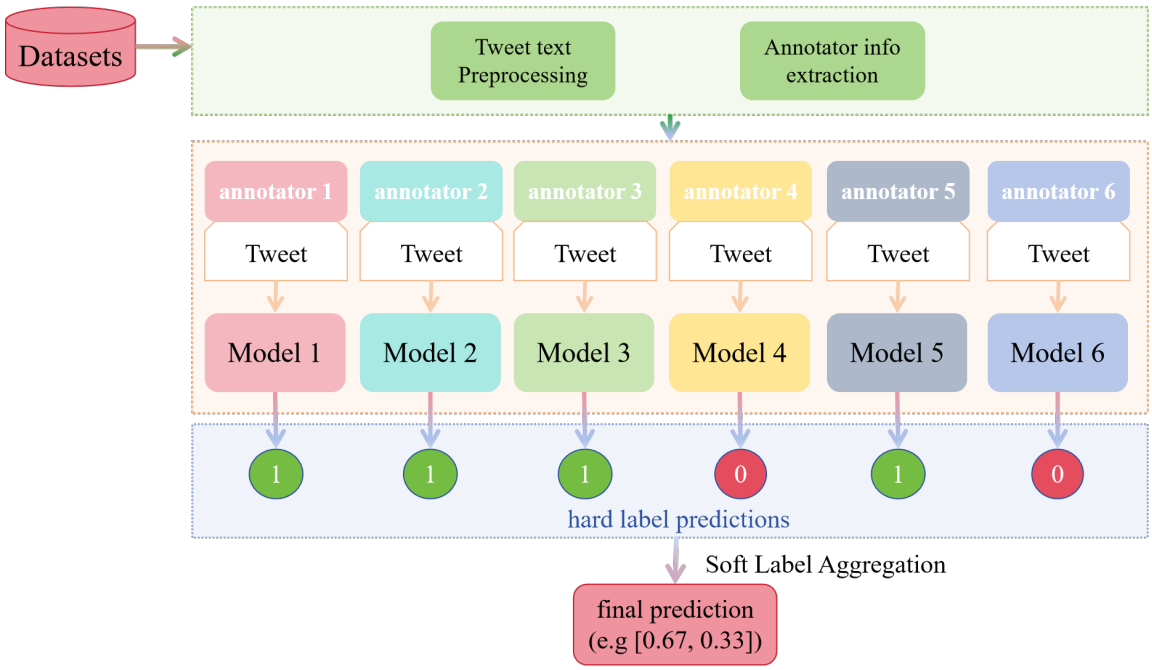

Figure 1: Annotator-aware multi-model ensemble architecture

4.1. Data Preprocessing

Given that the dataset comprises tweets from X, the textual content is inherently informal and often includes elements such as hashtags, user mentions, emojis, and URLs. These noisy components can hinder model performance if not properly normalized. To improve the quality and consistency of the input data, we applied a series of preprocessing steps aimed at reducing noise and linguistic variability—transforming raw tweets into normalized, standardized versions more suitable for model input. Below, we outline the key preprocessing operations applied during this stage:

1. Hashtags were stripped of the "#" symbol to retain only the core word.
2. User mentions were replaced with the placeholder token `user`.
3. URLs were substituted with the token `url`.
4. Emojis were converted into their textual representations using the `emoji` Python package, enabling standardized tokenization and reducing encoding inconsistencies.

To further increase data diversity and robustness, we employed the AEDA (An Easier Data Augmentation) technique [22]. AEDA introduces controlled perturbations by randomly inserting punctuation marks (e.g., ., ;, ?, :, !, ,) into sentences. In this way, the normalized tweets are transformed into augmented versions that enrich the training data without altering semantic content, thereby promoting model generalization.

4.2. Integrating Metadata

Each data sample is annotated by up to six annotators, with each annotator providing demographic metadata, including gender, age, ethnicity, education level, and country of residence. We split the dataset by annotator, resulting in six parallel subsets. Each subset contains the same tweet texts but is paired with metadata corresponding to a specific annotator, forming one-to-one tweet–annotator pairs.

To construct the input, we concatenate the original tweet with the annotator's metadata using the [SEP] token as a separator between fields. This format allows both the tweet content and the associated demographic information to be included in a single input sequence. Figure 2 illustrates an example of this formatting.

Original tweet: "it's a little old this video. but it is still very real for guys in the UK."

Annotator metadata:

- Gender: Female
- Age: 18-22
- Ethnicity: Black or African American
- Education: Bachelor's degree
- Country: South Africa

Processed input: "it's a little old this video. but it is still very real for guys in the UK. [SEP] gender:female [SEP] age:18-22 [SEP] ethnicity:Black or African American [SEP] education:Bachelor's degree [SEP] country:South Africa"

Figure 2: Example of Input Formatting with Annotator Metadata

This method integrates annotator information directly into the model input without requiring architectural modifications. Each tweet is thus associated with multiple instances, each reflecting a distinct annotator perspective.

4.3. Fine-tuning Models

To model the subjective perspectives of individual annotators, we fine-tune six model instances independently. While all sub-models share the same architecture and training configuration, each is trained on a distinct subset created by pairing tweets with the metadata of a specific annotator. Input sequences are formed by concatenating the tweet text with the annotator's demographic information, ensuring structural consistency across samples while enabling the model to capture perspective-specific patterns.

During training, we use the standard cross-entropy loss function as the optimization objective, defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{C} y_k \log(p_k)$$

where y_k denotes the one-hot encoded ground-truth label, and p_k is the predicted probability for class k. This loss function measures the discrepancy between prediction and truth, helping improve classification accuracy.

4.4. Post Processing

The final output is obtained by aggregating predictions from six independently fine-tuned models, each generating a discrete class label for a given input. We then aggregate these class predictions to construct a soft label—a probability distribution reflecting the relative frequency of each predicted class. To formalize this process, the probability $\hat{p}_k$ of class k is computed as:

$$\hat{p}_k = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(y^{(i)} = k)$$

where M denotes the number of models, $y^{(i)}$ is the predicted label from the i-th model, and $\mathbb{I}(\cdot)$ is the indicator function. This method captures both model consensus and uncertainty, aligning with the soft-soft evaluation protocol and effectively reflecting the subjective variation among annotators.

5. Experiments and Results

5.1. Experimental Setting

All experiments were conducted on a single NVIDIA A800 GPU (80GB memory). We utilized the HuggingFace transformers library for model training, with AdamW as the optimizer. The learning rate was set to 2e-5, and the weight decay was 0.1. Training was carried out for 8 epochs with a batch size of 16, and a linear warmup strategy was applied with 10% warmup ratio. The maximum input sequence length was set to 128. Mixed-precision training (fp16) was enabled to improve memory efficiency. Model evaluation was performed at the end of each epoch, and the checkpoint with the best validation performance was saved.

5.2. Main Results on Development Set

This section presents the results of our models during the development phase, under the soft-label evaluation setting. We assess three transformer-based architectures: ModernBERT-large, DeBERTaV3-large, and XLM-RoBERTa-large. Each model was tested across four configurations: (1) the baseline model, (2) baseline + data augmentation (DA), (3) baseline + annotator metadata integration (AMI), and (4) baseline + data augmentation (DA) + annotator metadata integration (AMI). To ensure a fair comparison, Task 2 models used ground-truth labels from Task 1 for the hierarchical structure, instead of predictions, allowing us to isolate the effect of architecture and training strategy.

Table 1
ICM-Soft scores on the development set for both tasks.

Model	Configuration	Task1 (ICM-Soft)	Task2 (ICM-Soft)
XLM-RoBERTa-large	Baseline	0.5042	-4.5517
	+DA	0.6098	-4.1429
	+AMI	0.6443	-3.9782
	+DA+AMI	0.7123	-3.4685
DeBERTaV3-large	Baseline	0.5187	-4.3891
	+DA	0.6332	-3.9874
	+AMI	0.6615	-3.8296
	+DA+AMI	0.7486	-3.3549
ModernBERT-large	Baseline	0.5619	-4.1581
	+DA	0.6781	-3.8316
	+AMI	0.6937	-3.5923
	+DA+AMI	0.7839	-3.2791

The results, as shown in Table 1, demonstrate that combining data preprocessing, data augmentation (DA), and annotator metadata integration (AMI) leads to consistent performance improvements across all models. In Task 1, XLM-RoBERTa-large improved from an ICM-Soft score of 0.5042 (baseline) to 0.7123 (+DA+AMI), and DeBERTaV3-large from 0.5187 to 0.7486. Similar gains were observed in Task 2, confirming the effectiveness of this strategy.

When examining the individual contributions of DA and AMI, we find that AMI yields stronger improvements when applied independently. For instance, with ModernBERT-large, the ICM-Soft score increased from 0.5619 (baseline) to 0.6937 using AMI, whereas DA achieved 0.6781—an approximate 13% relative improvement in favor of AMI. This suggests that annotator metadata contributes more directly to modeling subjectivity and improving the accuracy of learned label distributions.

Among all architectures, ModernBERT-large consistently achieved the best performance, with ICM-Soft scores of 0.7839 in Task 1 and -3.2791 in Task 2. Its strong adaptability to different annotator perspectives further demonstrates its superior ability to model subjective label distributions and to address the complex classification challenges posed by sexism detection. These results validate our choice of using it as the backbone for the EXIST 2025 task.

To further understand the model's behavior, we conducted an analysis of misclassified samples. We found that tweets with strong emotional tone or vulgar language were often misjudged as sexist, even when annotated as non-sexist. This suggests a bias in handling emotionally charged but gender-irrelevant content. Moreover, the model exhibited instability on samples with high annotator disagreement, reflecting the need for further improvement in handling highly subjective or semantically ambiguous texts.

5.3. Official Leaderboard Performance

Our final submission was based on the ModernBERT-large model, incorporating data augmentation and annotator metadata integration. In the final evaluation of EXIST 2025, our system ranked 7th out of 64 submissions for Task 1 in the Soft-Soft setting, and 20th out of 53 submissions for Task 2.

Table 2
Results for Task 1 Sexism Identification in Tweets

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	0	3.1141	1.0000	0.5770
EXIST2025-test_majority_class	61	-2.1991	0.1469	4.2166
EXIST2025-test_minority_class	66	-3.8158	0.0000	5.7521
mumule03_3	7	0.7135	0.6146	1.0876

Table 3
Results for Task 2 Author Intention Detection in Tweets

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	0	6.1178	1.0000	0.9354
EXIST2025-test_majority_class	34	-5.2028	0.0748	4.2291
EXIST2025-test_minority_class	55	-39.4948	0.0000	8.9579
mumule03_3	20	-3.5761	0.2077	3.1746

As illustrated in Tables 2 and 3, our proposed approach demonstrates competitive performance, particularly in Task 1, where it ranked 7th with an ICM-Soft score of 0.7135—substantially exceeding the baseline systems, which yielded scores of -2.1991 and -3.8158. While the system achieved a relatively lower rank of 20th in Task 2, it still outperformed the majority of baseline models in both ICM-Soft and cross-entropy metrics, reflecting its robustness and adaptability across tasks. These results underscore the effectiveness of the ModernBERT-large architecture in modeling annotator subjectivity. Overall, our method successfully captures the distributional nature of annotator disagreement under the Soft-Soft evaluation paradigm, delivering high-fidelity probabilistic outputs without introducing additional architectural complexity, thereby highlighting its practicality and scalability for real-world deployment.

6. Conclusion

This paper presents our approach for Tasks 1 and 2 of the EXIST 2025 shared task. We employed the ModernBERT-large model, enhanced through data augmentation and the integration of annotator demographic metadata to better model subjectivity in sexism detection. Our system demonstrated strong performance, ranking 7th in Task 1 under the Soft-Soft evaluation setting. These results highlight the effectiveness and practicality of our method in capturing annotator disagreement and addressing the nuanced nature of social bias in language.

Acknowledgments

This work is supported by the Quality Engineering Projects for Teaching Quality and Teaching Reform in Undergraduate Colleges and Universities of Guangdong Province (No.20251067).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-o3 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025 – learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2025.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, CEUR Workshop Proceedings, 2025.

[3] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in Human Behavior 52 (2015) 436–442. URL: https://www.sciencedirect.com/science/article/pii/S0747563215004641. doi:https://doi.org/10.1016/j.chb.2015.06.024.

[4] University College London, Social media algorithms amplify misogynistic content to teens, https://www.ucl.ac.uk/news/2024/feb/social-media-algorithms-amplify-misogynistic-content-teens, 2024. UCL News, published February 5, 2024.

[5] R. P. D. Díaz Redondo, A. F. Fernández Vilas, M. R. Ramos Merino, S. M. V. Valladares Rodríguez, S. T. Torres Guijarro, M. M. Hafez, Anti-sexism alert system: Identification of sexist comments on social media using ai techniques, Applied Sciences 13 (2023) 4341. URL: http://dx.doi.org/10.3390/app13074341. doi:10.3390/app13074341.

[6] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[7] N. Tahaei, S. Bergler, Analysis of annotator demographics in sexism detection, in: Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 376–383. URL: https://aclanthology.org/2024.gebnlp-1.24/. doi:10.18653/v1/2024.gebnlp-1.24.

[8] V. Prabhakaran, A. M. Davani, M. C. D'iaz, On releasing annotator-level labels and information in datasets, ArXiv abs/2110.05699 (2021). URL: https://api.semanticscholar.org/CorpusID:238634705.

[9] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[10] F. Wenzel, J. Snoek, D. Tran, R. Jenatton, Hyperparameter ensembles for robustness and uncertainty quantification, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 6514–6527. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/481fbfa59da2581098e841b7afc122f1-Paper.pdf.

[11] Y. Lin, C. Hsieh, H. Lee, Multitask multilingual learning with annotator demographics for sexism detection, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, Grenoble, France, 2024. URL: https://ceur-ws.org/Vol-3740/, eXIST 2024 Lab at CLEF 2024.

[12] V. Reyes-Meza, J. Gómez-Adame, H. J. Escalante, Hybrid systems for sexism detection: Combining gpt-4 and multilingual transformers, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, Grenoble, France, 2024. URL: https://ceur-ws.org/Vol-3740/paper-120.pdf, eXIST 2024 Lab at CLEF 2024.

[13] L. Bazikyan, M. Pérez, A. Gupta, C. Sánchez, et al., Soft label optimization with xlm-roberta for multilingual sexism detection, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, Grenoble, France, 2024. URL: https://ceur-ws.org/Vol-3740/paper-XYZ.pdf, eXIST 2024 Lab at CLEF 2024.

[14] A. Petrescu, Q. Rivas-Gervilla, A. Gutiérrez-Fandiño, L. Plaza, AlexPUPB at EXIST 2023: Identifying sexism in social networks, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023. URL: https://ceur-ws.org/Vol-3497/paper-088.pdf.

[15] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, ArXiv abs/2106.04681 (2021). URL: https://api.semanticscholar.org/CorpusID:235376976.

[16] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Thessaloniki, Greece, 2023, pp. 316–342. doi:10.1007/978-3-031-42448-9_23.

[17] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in tweets and memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Grenoble, France, 2024, pp. 93–117. doi:10.1007/978-3-031-71908-0_5.

[18] M. Siino, I. Tinnirello, Prompt engineering for identifying sexism using gpt mistral 7b, in: Working Notes of CLEF 2024, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1228–1236. URL: https://ceur-ws.org/Vol-3740/paper-115.pdf.

[19] J. Tavárez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, in: Working Notes of CLEF 2024, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1260–1273. URL: https://ceur-ws.org/Vol-3740/paper-118.pdf.

[20] A. F. M. de Paula, J. S. Culpepper, A. Moffat, S. P. Cherumanal, F. Scholer, J. Trippas, The effects of demographic instructions on llm personas, 2025. URL: https://arxiv.org/abs/2505.11795. doi:10.48550/arXiv.2505.11795. arXiv:2505.11795, accepted at SIGIR 2025, Padua, Italy.

[21] A. Jiang, N. Vitsakis, T. Dinkar, G. Abercrombie, I. Konstas, Re-examining sexism and misogyny classification with annotator attitudes, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15103–15125. URL: https://aclanthology.org/2024.findings-emnlp.887/. doi:10.18653/v1/2024.findings-emnlp.887.

[22] A. Karimi, L. Rossi, A. Prati, Aeda: An easier data augmentation technique for text classification, 2021. URL: https://arxiv.org/abs/2108.13230. arXiv:2108.13230.